ビッグデータの活用事例と求められるデータ・サイエンティストとは

総合情報基盤センター　教授　高井 正三

“ビッグデータ”という言葉が出始めたのは 2010 年末頃からで，2013 年 5 月 20 日に発行された「ビッグデータの正体（V.M=ショーンベルガー&K.クキエ著，講談社刊）[3]」を契機に，新聞・雑誌で頻繁に登場するようになった．2014.10.16 版日経コンピュータの特集 5 部「ビッグデータ，夜明け前」で「勤務先では，ビッグデータ活用に本腰を入れていますか」の調査に解答したユーザ企業 1,752 社中で，84%が「本腰を入れていない」と回答し，12%が「本腰を入れているが効果が出ていない」，「本腰を入れて効果が出ている」企業は 3%であった．私たちが身近に体験している Amazon.com の「よく一緒に購入されている商品」「この商品を買った人はこんな商品も買っています」と表示して，更なる購買意欲を刺激してくる表示こそは，最たる活用事例だが，本稿では，今後のビッグデータの益々の活用を願い，「求められるデータ・サイエンティスト」を提案したいと思う．話題の IoT，機械学習を始め，大学での IR（Institutional Research）戦略などに，是非本提案を活かしてもらいたい．

1．我が国におけるビッグデータの活用事例

1.1　コマツの KOMTRAX（コムトラックス）

2011 年 4 月 8 日に発行された「ダントツ経営（著者：坂根正弘＝当時コマツ会長＝現相談役，日本経済新聞社刊）」という著書の第 1 章で“コムトラックスで市場を「見える化」する”が紹介されている．KOMTRAX はコマツの建設機械に標準装備されている，稼働状況を遠隔監視できる ICT システムであり，1999 年から稼働し，世界各地で稼働するコマツの建機に取り付けられた GPS や各種センサーから，現在の位置，稼働時間，稼働状況，燃料の残量，消耗品の交換時期などのデータを，通信衛星と Internet 経由でコマツのデータ・センターのサーバーに送信されるシステムである．

ビッグデータ（BD：Big Data）時代の先駆けであり，我が国におけるデータフィケーション Datafication（「すべてのもの」をビジネスに活用できるようデジタル・データ化すること）を具現化した最初の例である．KOMTRAX で，世界各地の販売代理店や顧客はコマツのサーバーにアクセスして，自分の地域のデータや，顧客が自信のデータを確認できるため，GPS により，どの地域で建機の稼働時間が増加し，どの地域で減少しているかも把握できるので，需要動向を予測し，在庫や生産量を適切にコントロールできるようになり，消耗品の交換など，建機の予防保守も可能になった．2012 年 3 月末時点で，全世界 70 か国で，26 万台の建機で稼働中であると言うが，本当はリース料金を支払わない顧客の機械を遠隔ロックすることもできるようだ[4].

1.2　Amazon.com（アマゾンドットコム）

身近な事例としては Amazon.com で，商品を検索した結果，追加情報として表示される購買を刺激する情報である（図 1，図 2）.

図 1　「ビッグデータの正体」を検索

図 2　本の後に提供される，購買を刺激する情報

1.3　国立科学博物館

2014.7.24 発行の日経コンピュータの特集「格差広げるビッグデータ 100」の最初の活用事例（p.31）として，国立科学博物館では“人流”をセンサーで全記録を収集し，乃村工藝社，日立製作所と共同で，この記録を解析し，見学ルートの改善，子供と大人の展示解説を分けるなど，効果的な見学ルートの設計に役立てている（図 3）[1].

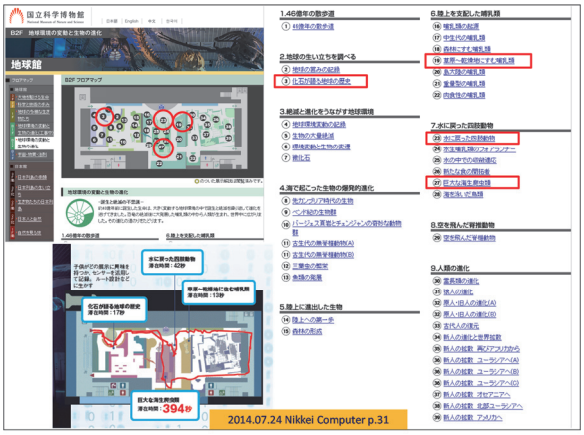

図３　国立科学博物館でのビッグデータ活用事例

1.4　データを上手に利用する企業

（1）リクルート

リクルートは，Hadoop の徹底活用でデータ分析に対する意識改革に成功し，「SUUMO」「ゼクシィ」「じゃらん」「ホットペッパー」などで活用されている．中古車情報サイト「カーセンサーNet」では，割引チケット共同購入サイト「ポンパレ」など，企業と人を結び付ける多彩なサイトを運営し，「ホットペッパー」では，1 週間分のアクセス・ログを処理するのがやっとで，一部の会員 8 万人に Recommend Mail を送付していたが，Hadoop で，1 年半のログを処理し，20 万人に Recommend Mail を送付できるようになった[4].

（2）GREE

GREE では，急成長の原動力となるデータ駆動型アプローチで，2011 年第 4 四半期で DeNA を抜いた．「1 個人のセンスよりも数千万人のデータを信じる」として，GREE Analytics という Data Mining Tool を独自開発し，ユーザーの登録日，登録経路，利用状況，各イベントの参加率，プレイ率，アイテム別売上げ，ゲーム進捗状況，継続率などのユーザー動向データが，時間単位で把握できるようになった[4].

（3）日本マクドナルド

最近何かと問題の多い日本マクドナルドは，携帯電話サイトの「トクするケータイサイト」なる One to One マーケティング・サイトを 2003 年 7 月に起ち上げ，2011 年 3 月には，おサイフケータイ対応携帯電話向け「かざすクーポン」を開始した，同社の顧客 1 人ひとりの購買履歴を詳細に分析し，購買パターンに応じて，1 人ひとり内容の異なる割引クーポンを配信し，サービスしている[4].

1.5　トヨタとホンダの活用事例

2014.10.10 の日経新聞記事によれば，トヨタが 2014 年 6 月に発表したテレマティクス・サービス「T-Connect」は，カー・ナビゲーションで設定した走行ルート上で渋滞発生を予測すると，それを回避するルートを運転者に勧める．一方，ホンダは自社のカー・ナビゲーション・システム「インターナビ」から匿名で自動車の動作情報を収集し，急ブレーキ多発地点を割り出し，交通安全情報を提供している Web サイト「セーフティ・マップ」に掲載している．

1.6　ビッグデータ最新の活用事例

2015.1.1 版の日経新聞第 2 部記事「デジタルが運ぶ未来」によるビッグデータ活用事例．

（1）IHI のガスタービン運用支援システム

米 GE の風車発電での BD 活用事例の後に，IHI は 2013 年末，国内外に納めた 136 基の発電用ガスタービン・システムを一元的に運転支援・管理する，Global Monitoring & Technical Service Center（i-MOTS）を設立，ガスタービンにセンサーを取り付け，タービンの回転数や振動，機器温度など，200～300 種類のデータを，1 分間隔で取得出来るようになっており，障害の予兆を察知すると，IHI の担当者に警報を鳴らして対処するという．

（2）東工大とアステラス製薬

東工大の秋山泰教授は，2013 年，東京大学やアステラス製薬と共同で，熱帯感染症について世界各地でまとめられた論文を統合したデータベース「iNTRODB」を構築，関嶋准教授とアステラスなどは，これを活用しリーシュマニア病，シャーガス病，アフリカ睡眠病の病気に効果のある治療薬の開発を目指しているという．先ず，市販されている 2,000 万種の化合物の中から，効果の可能性のある 500 万種を選び出し，その上で，世界中の研究論文を基礎データとして，同大のスーパー・コンピューター「TSUBAME」を用いて，実際に寄生原虫のタンパク質に結合するかどうかなどを計算し，最終的に化合物を 1,000 種に絞った．計算で可能性が認められた物質をアステラスが実験し，20 種の医薬品候補が得られているという．現在はデング熱についても同様の作業を実施中であると言っている．

1.7　経済産業省の情報通信白書

平成 26 年版情報通信白書では，注目のビッグデータ活用事例として以下を挙げている．

◆製造業・・・マツダ（株）
◆農業・・・本川牧場，◆水産業・・・（株）グリーン＆ライフイノベーション
◆サービス業・・・（株）あきんどスシロー
◆運輸業・・・イーグルバス（株）
広告業・・・　（株）マイクロアド

2．海外におけるビッグデータの活用事例

2.1　米サンタクルーズ Santa Cruz 市警

2011 年 7 月．米カリフォルニア州サンタクルーズ市で不思議な現象が起こった．犯罪が発生する前に，犯罪現場に警察官が現れるようになったのである．それから 3 年，同市では実際に犯罪発生件数が 17%も減少したという．これは，プレディクティブ・ポリシング(Predictive Policing=予測警備)という，犯罪予測システムを導入した結果であるという．

今までの犯罪データを分析した結果，

◆Repeat Victimization（一度被害にあった場所で 2 週間以内に被害が再発するという傾向）
◆Near Repeats（犯罪が発生した近郊で犯罪が再発しやすいという傾向）

から，サンタクルーズ市警は 2011 年 7 月に，モラー博士らが開発した予測モデルを搭載した犯罪予測システム「PredPol」を導入した．
URL=http://itpro.nikkeibp.co.jp/atcl/watcher/14/334361/080100020/?SS=imgview&FD=1124500606&ST=bigdata（Nikkei ITPro）

写真1●「プレディクティブ・ポリシング」を導入したサンタクルーズ市警

写真2●サンタクルーズ市警で副署長（Deputy Chief of Police）を務めるスティーブ・クラーク（Steve Clark）氏

写真3●サンタクルーズ市警のパトカー車内

写真 1,2,3　サンタクルーズ市警（日経 ITPro）

犯罪予測システム「PredPol」では，「車上荒らし（Vehicle Burglary）」「住居への強盗（Burglary）」「自動車窃盗（Auto Theft）」「拳銃やナイフを使った犯罪（DW Assault，DW は Deadly Weapon の略 ）」「拳銃などを使わない暴行（Battery）」といった犯罪が，昨日どこで発生し，これからどこで発生しそうか地図上に表示する（写真 1～5）．

写真4●カリフォルニア大学ロサンゼルス校（UCLA）のジェフ・ブランティンガム（Jeff Brantingham）博士が開発した犯罪予測モデル

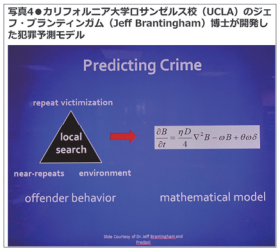

写真5●「PredPol」による犯罪発生予測画面

写真 4,5　PREDPOL 予測モデル／画面とサイト

（**参照**：2014.7.24，日経コンピュータの特集「格差広げるビッグデータ 100」の第 17 位の活用事例，p.35／前掲 Nikkei IT Pro）

2.2　米ビッグデータのバリュー・チェーン

米国でビッグデータを使って Value Chain（価値連鎖＝価値を高めていく）の事例として，データ型，スキル型，アイデア型の 3 つの企業タイプがあると，前著第 7 章で記述しているのでるので，それを観てみよう．

（1)航空券予約ネットワークを運営する ITA ソフトウェア（データ型の事例）

航空運賃予測サービスのフェアキャストにデータを提供しているが，自社では分析作業をしていない．フェアキャスト社は航空機のチケットをいつ購入したらいいのか＝安く買えるかを予測する会社である．

同社の創業者エレン・エツィオーニは，数カ月前にチケットを入手していたにも関わらず，他の乗客よりも高く買っていた．この悔しさをバネに VC（ベンチャーキャピタル）から資金を調達. すべての路線の全フライト，全座席を 1 年間追跡し，チケットの価格を予測できるようになった．エツィオーニは予測精度を高めるために，旅行業界向けのフライト予約データベースに触手を伸ばした．2008 年頃から，ホテルやコンサートのチケット，中古車などにもこの手法を利用しようと考え始めた．それを評価した米マイクロソフトが同社を 1 億ドルで買収した．

（2）Master Card（データ型・スキル型）

クレジット・カードの Master Card は自社でデータを分析している．同社のカード会員は 210 カ国に 15 億人おり，Master Card Advisers と呼ばれる部門が，650 億件の取引データを集めて分析し，ビジネスと消費者のトレンドを予測する．このトレンド情報を外部に販売している．

（3）アクセンチュア（スキル型）

スキル型とはデータベース・スペシャリスト企業で，具体的には複雑な分析を実施するノウハウや技術のある企業である．

アクセンチュアは，様々な業界から委託を受けて，最先端の無線センサーでデータを収集し，分析している．ミズーリ州セントルイスの市営バスに無線センサーと取り付け，エンジンをモニタリングし，故障発生の予測や最適な定期保守の判断に役立てた．この結果車両保有コスト 10%を削減，バス 1 台当たり $1,000 を削減することができた．

（4）Microsoft Research（スキル型）

Washington DC にあるメドスター・ワシントン医療センターでは，再入院や感染症を抑えるため，Microsoft Research （MR）に委託して，匿名化した診療記録数年分を分析した．診療記録には，患者の属性情報，検診結果，診断，治療などが記載されている．使用したソフトウェアは MS の「アマルガ Amalga」で，分析の結果，驚くべき相関関係がいくつか見つかった．退院後 1 ヶ月以内に再入院する可能性が高まった条件を一覧にまとめた．その分析から，

・鬱血性心不全の患者は再入院しやすく，再入院時は治療も難しくなるが，予想外な兆候が見つかった．

・「憂鬱感」など心痛らしき言葉が含まれていた場合，退院から 1 か月以内に再入院する確率が著しく高まることが分かった．

（5）Flight Caster.com（アイデア型）

Bradford Cross は 2009 年 8 月，友人等と「フライト・キャスター・ドットコム Flight Caster.com」を立ち上げた．すでに公開されている過去 10 年の全フライトを気象データと組み合わせ、米国内のフライトの遅延予測情報を提供している．その後 Cross はニュース・サービスに目をつけ，プリズマティック Prismatic というベンチャー企業を起ち上げ，テキスト解析，ユーザーの好み，SNS 関連の人気など，ビッグデータの解析から，Web コンテンツを集めてランク付けをしている．

（6）ビッグデータ思考の企業や個人の例

・交通量分析のインリックス Inrix

・eBay…毎日 50TB のデータが生成

・Zynga…ゲーム会社の皮を被った分析会社
・Centrica…スマートメーター（通信機能を備えた電力メーター）導入によりエネルギー消費パターンを分析
・Catalina Marketing…レジ･クーポンで顧客の購買行動をデザイン

３．ビッグデータと３つの大変化

3.1　ビッグデータ以前

既にスーパー・マーケットの Point Card や POS（Point of Sales）端末で，ユーザーの層と天候，曜日，時間帯と購買情報の関連が分析されて，広告の作成や商品の仕入れ，陳列に活用されている．ビッグデータ以前はソーシャル・メディア・リスニング Social Media Listening と言われ，2011 年，富山県内ではアルミ製品の三協立山（株）が既にマーケティングに活用している．Social Media Listening とは， Facebook，Twitter 上で展開される企業や商品に関する生活者の口コミ情報を収集／分析することで，Facebook 以上に情報が入手しやすい Twitter がターゲットになっている．Twitter の情報はフリーの分析サイトや,「見える化エンジン」を提供しているプラスアルファ・コンサルティング，Facebook も同様の Buzz Finder や True Teller の他, Salesforce.com の Radian6 などのテキスト・マイニング分析システムによって，つぶやき情報，アカウント情報，アクセス解析情報などから分析がなされ，自社のアカウント／ブランディング／キャンペーン／競合分析，関連ワードや発言者分析などが行われ，企業の商品やサービスの戦略に利用されていた. 企業の Facebook 活用事例として，米国ではナイキやコカコーラ，スターバックスが，国内では Satisfaction Guaranteed，ユニクロ，無印良品，楽天市場などが「ファンページ」を開設し，その情報を分析して，マーケティングを行っている．

3.2　ビッグデータとは

（1）総務省情報通信白書（H26 年度版）でのビッグデータの定義

白書では，鈴木良介著「ビッグデータビジネスの時代」を参照し，ビッグデータとは，「事業に役立つ知見を導出するためのデータ」と定義し，ビッグデータ・ビジネスを,「ビッグデータを用いて社会・経済の問題解決や，業務の付加価値向上を行う，あるいは支援する事業」と定義している．

（2）ビッグデータ関連図書のベース著書「ビッグデータの正体」では p.18 から，「小規模ではなしえないことを, 大きな規模で実行し，新たな知の抽出や価値の創出によって, 市場，組織，さらには市民と政府の関係などを変えること.」，それがビッグデータである．

（3）2012 年 2 月発行の The Economist 誌特集“The data deluge「データ大洪水」”が契機となって,「ビッグデータとは, 既存の一般的な技術（RDBMS：関係型データベース管理システムなど）では管理するのが困難な大量のデータ群である」と定義され，ビッグデータの特性は 3V（Volume，Velocity, Variety：量（＝データ量），速度（＝入出力データの速度），多様さ（＝データの型，データ発生源，データの範囲））で示される．

（4）Big Data の定義（Gartner）

Gartner は US 版 Wikipedia で次のように定義している．（日本版はこの直訳を掲載）

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data.

Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.

In a 2001 research report and related lectures, META Group (now Gartner)

analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources).

Gartner, and now much of the industry, continue to use this "3Vs" model for describing big data.

In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, in－sight discovery and process optimization."

3.4　ビッグデータの量

南カリフォルニア大学コミュニケーション学部のマーティン・ヒルバート教授は, 書籍, 絵画, メール, 写真, 音楽, 動画（Analog/Digital), テレビゲーム, 電話通話, カーナビ・システム, 放送メディアの視聴率から算出し, 2007 年 300EB（Exa Bytes, 10^{18} Bytes）としている[3].

日本アイ・ビー・エムでは,

2009 年の年間, 0.8ZB, 毎日 2.5EB 生成.

2011 年の年間, 1.8ZB（Zetta Bytes）.

2020 年の年間, 35ZB（予測）.

（Zetta Bytes＝10^{21} Bytes）としている[5].

3.5　ビッグデータ「3 つの大変化」

前著「ビッグデータの正体」によると, 3 つの大変化とは以下の通りである[3].

（1）第 1 の変化「すべてのデータを扱う」「N=全部」の世界

◆無作為抽出という革命

無作為抽出した 1,100 人の標本があれば 97%以上の精度で, 母集団の動向を言い当てることができる. 400 人無作為データでは, 95%の確率で, 1 万人から, 10 万人, 100 万人, 1,000 万人, 1 億人の意見が分かる.

◆標本作成の失敗例

1936, 当時存在した有力週刊誌「Reader Digest」が, 大統領選を前に有権者 200 万人を対象に調査を実施, 共和党候補の圧勝を予測したが, これが大外れで, Franklin D. Roosevelt が 523 対 8 で大統領選に圧勝した.

原因は無作為性が甘かった. 同誌は購読者リストと電話帳により調査対象者を選んだのだが, 当時としては電話を所有しているのは裕福者で, 共和党支持者が多かった.

◆八百長試合を探せ

角界を揺るがす八百長疑惑。日本相撲協会の放駒理事長は 2011 年 2 月 2 日の会見で「過去には一切なかった」と述べたが, シカゴ大学のスティーブン・レビット教授等は, 1989 年から 2000 年までの, 十両以上の力士 281 人の取組 32,000 回以上を調べた. その結果, 千秋楽に 7 勝 7 敗の力士が 8 勝 6 敗の力士と対戦した際の, 勝率の「からくり」を過去の対戦結果から出した計算では, 7 勝 7 敗の力士の勝率は 48.7%だが, 7 勝 7 敗で迎えた力士の千秋楽での勝率は 79.6%にもなった.

この確率は, 次の場所で両者とも勝ち越し問題が生じない場合, 7 勝 7 敗の力士の勝率は 40%にダウン. その次の場所では約 50%と, 元の勝率に近づくという.

レビット教授と同僚のマーク・ダガン教授は, 過去 11 年分, 延べ 6 万 4000 番の取組データを基に異常を探し出した. 目論見は当たった. 確かに八百長試合らしき動きがあったが, 誰も注目しないような取組だった. この奇想天外な研究論文は, 学術誌の「American Economic Review」に掲載され, 後に「Freakonimics（邦訳『やばい経済学』共著, 東洋経済新報社）」として出版され, ベストセラーになっている.

（2）第 2 の変化「精度は重要ではない」量は質を凌駕する

「乱暴な方が正確になる」時代

◆文法チェッカー（Microsoft）

2000 年 MS Research のミシェル・バンコとエリック・ブリルが MS Word の文法チェッカーの改良を模索していた.

既存のアルゴリズムで，データ量を増やすことを確かめる．通常は 100 万語のコーパス（Corpus:実際の文例 DB）だが，2 人は 4 つのアルゴリズムを用意し，1000 万語，1 億語，10 億語でトライした．50 万語で最低の成績だった単純なアルゴリズムでは，10 億語で，文法ミスを見つけ出して修正する正答率が 75%から 95%以上に跳ね上がった．

最高のアルゴリズムでも正答率は 86%から 94%に改善されただけだった．

◆Google は 1 兆語で，Google 翻訳に挑む．

2006 年，Google が誇る 1 兆語 Corpus に収録されている英語センテンスは，品質は怪しいが，950 億語を達成し，翻訳サービスは，精度も高く，最もうまくいっている．

2012 年半ばには，対象言語が 60 に拡大，14 言語では音声入力でも，円滑な翻訳が可能になった．

◆機械翻訳（IBM）

1954 年，IBM701 で 250 語の言葉のペアと 6 つの文法ルールを登録し，ロシア語の 60 フレーズを英語に，円滑に翻訳した．

1990 年代後半，IBM の「キャンディード」プロジェクトでは，英語とフランス語で発行されているカナダ議会の議事録から 10 年分に及ぶ翻訳，およそ 300 万センテンスを利用して，機械翻訳をおこなった．成果は今ひとつだった．

◆量は質を凌駕する

ビッグデータの世界に足を踏み入れるためには，「正確=メリット」という考え方を改める必要がある．

◆ビリオンプライス・プロジェクト

米労働統計局は，消費者物価指数の算出に，全米 90 都市の小売店や企業を対象に，数百人もの職員が日々，電話，ファクス，直接訪問による聞き取り調査を実施した．

トマトの料金からタクシー料金まで，8 万点の価格を，年間 2 億 5 千万ドル（250 億円）を使って，数週間かけて報告書としてまとめていた．

MIT の経済学アルベルト・カバロ教授とロベルト・リゴボン教授はビッグデータを使って物価調査を実施．Web 上のデータを自動的に集めるソフトを駆使し，毎日 50 万点の価格を収集する．

このビッグデータに，ある分析を加えた結果，2008 年 9 月のリーマンショック後のデフレ兆候を見抜いた．

（3）第 3 の変化「因果から相関の世界へ」答えが分かれば，理由はいらない

◆書評家を敗北させたアマゾン

Washington 大学大学院で人工知能を研究していたグレッグ・リンデン Greg Linden（24）は，1997 年に休学し，オープンから 2 年の Amazon.com で働くことにした．

同社の Web site に，当時の競争力の源泉でもあった「アマゾンの声」という書籍紹介コーナーがあった．

同社 CEO のジェフ・ベゾスがある有望なアイデアの実験に乗り出す．「個々の顧客の購入履歴や好みのデータに基づいて書籍を推薦する仕組み」や，顧客の膨大なデータ（「最後まで迷ったが，購入に至らなかった書籍」「どれくらいの時間チェックしていたか」「一緒に購入したのはどの書籍か」）を蓄積した．このデータを従来の方法「標本データを分析し，顧客全体の共通項を探る」で加工していた．その結果，「前回の購入書と大差ない書籍を延々と紹介し続けた．客にしてみれば，はた迷惑な店員につきまとわれながら買い物をしているようなものだった」（当時の書評委員：ジェームズ・マーカス）

Greg Linden は，顧客全体の買い物内容から共通項を探る機能は，商品推薦システムに不要だと気付き，重要なのは，一見関係なさそうな商品同士の相関関係を見つけることだった．Linden 等は，「商品間」の強調フィルタリング技術で特許申請し，この手法に切り替えたことが転換点となった．

相関関係の計算は予め済ませておけるので，お勧め商品は即座に表示でき，汎用性も高く，

商品カテゴリーにまたがるお勧めも可能になった.

次は提示する内容．専属の書評委員による書評か，それともコンピューターがはじき出した顧客別のお勧めやベストセラー・リストか．書評委員の言葉を信じるか，蓄積されたクリックの“声”を信じるか.

Linden は，この両者から販売に繋がったケースを比較．差は歴然で，コンピューターのデータから導出したコンテンツが 100 倍も大きな売り上げを生み出していた

百田尚樹を読んだ後に，なぜ jQuery の本を買いたいと思ったのか，コンピューターは知る由もない．それは重要ではなく，ともかく売れたことが事実である（筆者の例）.

やがて，人間の手による書評がオンラインで公開されるたびに，書評委員らに正確な売り上げデータが突きつけられた．そしてついに書評チームは解散を余儀なくされた.

Linden は，「書評チームが負けたことはとても残念だった．しかしデータは嘘をつかない. コストも非常に高かった.」と言っている.

現在，Amazon.com の売上げ全体の 1／3 は，この「おすすめ」とパーソナル化のシステムから生み出されているという．Linden の技術は, Online 販売の世界に革命をもたらしたのである[3].

◆ネットフリックス Netflix

Online DVD レンタルのネットフリックス Netflix, Inc．では，新規受注の 3/4 が推奨作品である.

◆ビッグデータの先駆者―ウォルマート

ハリケーンの到来が近づくと，懐中電灯と「ポップターツ」の売上げが増加する，という事実が判明した．そこでハリケーン対策用品コーナーに「ポップターツ」も大量に陳列したところ，大いに売上げを増大した.

◆主役に躍り出た「相関分析」

購入品目から女性客の妊娠まで予測した例から，各方面に応用される「予測分析」で，因果関係はそこまで重要なのか.「オレンジ色のクルマはなぜ欠陥が少ないのか?」あなたは分かりますか？事実なのである．理由なんかないのである.

理論は終焉するのか，という問いに，ペタバイトのデータがあれば,「相関で十分」と言えるのである.

3.6　データフィケーション Datafication

「すべてのもの」がデータ化され，ビジネスになる時代が到来した.

◆「座り方」データが有望なビジネスに変身

産業技術大学院大学の越水重臣准教授は，人間の臀部の形状を科学的に捉える研究に取り組んでいる. 着座したときの尻の形, 姿勢, 重量分布を数値化・集計することで，座り方自体が情報になるという．自動車のシートに 360 個の圧力センサーを取り付け，着座時の圧力を 256 段階で測定し，臀部をデータ化している.

この得られたデータは 1 人ひとり違うことが分かり，実験では，数人の被験者を 98%の精度で識別できた.

この技術を，自動車盗難防止システムの開発に応用し，登録ドライバー以外が運転席に座ると，パスワードが求められ，認証に失敗するとエンジンはかからないようにする.

この技術の応用は，運転時のドライバーの姿勢も記録されるので，交通事故を防ぐための自動ブレーキかけや, ひき逃げなどの同定, 危険防止の警告鳴らしなどに使えるという.

◆位置もデータに変わる

人間の行動を逐一記録するアプリケーションが登場している.

Google の Street View は, 街の写真を撮影する際に, 近隣から電波が漏れ出ている WiFi ルーター情報も収集している.

iPhone には位置情報と WiFi データを取得して Apple に送り込む機能が入っていた（Android や MS の携帯向けも同様）.

米大手運送会社 UPS は保有車両にセンサー, 無線モジュール, GPS を取り付けている. このシステムに知恵や洞察力が生まれる.

・エンジン故障を未然に予測.
・配送遅延の有無やドライバーの状況チェック
・過去の輸送・配送データから最も効率的な最適ルートの作成で，2011 年に，走行距離 4,800 万㎞，ガソリン 600 万リットル，3 万トンの CO_2 削減に成功.
・交差点での右左折の少ないルートをアルゴリズムで同定し，安全性や業務効率を向上.

◆その他の Datafication

・「Foursqure」というアプリでは，指定された場所を訪れた印として「check-in」ボタンを押すと Point がもらえる．Foursqure 側には客を運んだ謝礼として，各種ポイント・サービスやレストラン案内サービスなど位置情報関連サービスから報酬が支払われる仕組み.
・Amazon.com でのショッピング，クリック，カスタマーレビュー
・Google の様々なサービスでのクリック
・Facebook での投稿や「いいね」の他，人間関係をグラフ化する「Social Graph」
・Twitter での tweet や Retweet から「心の動き」をデータ化
・LinkedIn でも，・・・Google+でも，Tumblr, Pinterest でも，・・・

4．ビッグデータのマイナス面

4.1　ビッグデータのマイナス点項目

ビッグデータのマイナス点は以下に挙げるように多々あるので，その項目を挙げる.

・Amazon.com・・・ショッピングの好み
・Google・・・Web Site 閲覧の癖
・Twitter・・・心の動き
・Facebook・・・心の動き＋交友関係
・SmartPhone・・・通話相手＋すぐ近くにいる人物
・街角の監視カメラ・・・移動状況，プライバシーの保護が困難になる．プライバシーへの脅威を生み出す．データ独裁の犠牲者になるリスク
・プライバシーの麻痺
・匿名化されたデータでも同定は可能
・データの独裁が可能

4.2　プライバシー保護のために使われてきた 3 大対策

・個別の告知と同意
・データ利用拒否を本人が通知できる精度 OptOut
・匿名化

4.3　根底から変わる捜査のあり方

・予防型犯罪捜査
・映画「Mynority Report」の例

５．求められるデータ・サイエンティスト

5.1　データ・サイエンティストとは

著書「“ビューティフルデータ Beautiful Data”，Toby Segaran, Jeff Hammerbacher 編，堀内，真鍋，苅谷，小俣，篠崎共訳，オライリー・ジャパン， 2011.2.28，ISBN978-4-87311-1489-7，￥3,400+TAX」では，次のように肩書きを作ったようだ.

Facebook では，ビジネス・アナリスト，統計学者，エンジニア，リサーチ・サイエンティストといった従来の肩書きは，私たちのチームにとってまったく魅力的なものではなかった. 各役割の作業負荷は多種多様である.

ある日の，あるメンバーの行動は，(1) 多段階の処理パイプラインを Python（言語）で書き，(2) 仮説検定を設計し，(3) 統計ソフトウェア R を用いてデータ・サンプルの回帰分析を行い，(4) Hadoop で大量のデータを扱う製品やサービスのアルゴリズムを設計して実装し，(5) 分析結果を明瞭かつ簡潔な方法で，組織の他のメンバーと話し合う，といった感じだ.

このように数多くの仕事をこなすのに必要なスキル一式を著すために，私たちは “Data Scientist（DS）”という肩書きを作りだした.

5.2　データ・サイエンティストに求められるスキル（skill：技能）

以下のような skill が必要不可欠である.

(1) Computer Science…Hadoop や Mahout などの大規模並列処理技術や機械学習，Database, RDBMS と SQL, Python/PHP な

どの Script 言語，修士号/博士号または同等の職務に 4 年以上の経験.

（2）数学，統計，データマイニング…統計パッケージ SPSS, SAS などの技術の他, OSS プログラミング言語 R の技能

（3）データの可視化…SAS，MATLAB，R，Infographics の技能

5.3 Facebook の Data Scientist に対する求人票の内容

［職務内容］

（1）重要なプロダクトの課題を同定し，対処するために，Product Engneering Team と密接に連携して職務にあたる.

（2）データに対して，適切な統計テクニックを適用し，課題解決を図る.

（3）結論を Product Manager と Engineer に伝える.

（4）新規データの収集と既存のデータソースの改良を推進する.

（5）Product の実験結果を分析・解明する計測・実験方法の Best Practice を開発し，Product Engneering Team に伝える.

［資質］

（1）コミュニケーション能力.（2）起業家精神.（3）好奇心

5.4 データサイエンティスト協会が求めるデータサイエンティスト(DS)のミッション，スキルセット，定義，スキルレベル[8]

2015 年 1 月 5 日付け日経産業新聞（p.7）の囲み記事「データサイエンティスト　スキル定義　育成の基準に」と言うタイトルで，一般社団法人データサイエンティスト協会（東京・港区，代表理事：草野隆史）が，DS のスキル定義を発表したと報道している．同協会のホームページから，そのミッション，スキルセット，定義，スキルレベルとは，

（1）DS のミッション Mission

人間を数字入力や情報処理の作業から開放するプロフェッショナル人材であり,「データの持つ力を解き放つ」こと.

（2）DS に求められる Skill Sets

1)ビジネス力（business problem solving）：課題背景を理解した上で、ビジネス課題を整理し、解決する力

2)データサイエンス力（data science）：情報処理、人工知能、統計学などの情報科学系の知恵を理解し、使う力

3)データエンジニアリング力（data engineering）：データサイエンスを意味のある形に使えるようにし、実装、運用できるようにする力（図 4）

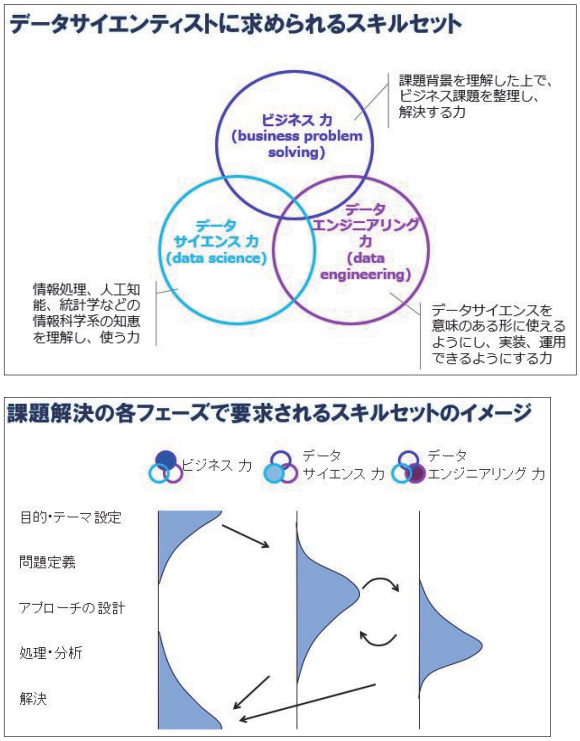

図４　データサイエンティスト協会のスキルセット

「データサイエンティストとは，データサイエンス力，データエンジニアリング力をベースにデータから価値を創出し，ビジネス課題に答えを出すプロフェッショナル」

（4）DS のスキルレベル Skill Level

1)業界を代表するレベル：

Senior Data Scientist

2)棟梁レベル：(full) Data Scientist

3)独り立ちレベル：

Associate Data Scientist

4)見習いレベル：Assistant Data Scientist

［注］「Senior Data Scientist　（業界を代表するレベル）は，一人である必要はないと考

えます．一人で現実的に全て持てる多くの場合の目標点が、(full) Data Scientist（棟梁レベル）という見立てです．全体をコーディネートし，俯瞰できる人は必要ですが，加えて個別のスキルセットで秀でた人とのチームを作り，推進することも現実的には多いためです.」と注意書きしている．

６．ビッグデータが未来を変える

6.1　人工知能とディープラーニング

2014.10.2 号版 NIKKEI COMPUTER[9] の特集第 1 部「人工知能を制する者が勝つ」と第 2 部「ディープラーニングの衝撃」で，人工知能（AI：Artificial Intelligence）を制する者がビッグデータを制し，更にビジネスを制する．その鍵となっているのがディープラーニング（Deep Learning：深層学習）であるという.「機械学習」即ち，テキストや画像，音声といったデータから意味を認識するためのパターンやルールを，コンピューターが自動的に見つけ出す技術が，人工知能のレベルを驚異的に引き上げている．

デンソーIT ラボラトリーの画像認識システムなど，今注目を集めているのは，脳の仕組みを模した「Deep Neural Network」というシステムを使用する機械学習であるという．

Google が開発している自動運転システム，Apple 社の音声アシスタント機能「Siri」など，人間では扱いきれない大量の「ビッグデータ」から，人間とほぼ同じレベルで意味や知識を獲得できるようになるからである．

6.2　人工知能マシン Watson と Twitter

米 IBM は 2014 年 10 月 29 日（米国時間），米ツイッターとビジネス向けビッグデータ解析で提携すると発表し，Twitter 上のつぶやきを分析し，活用する業務アプリケーションを，銀行や消費財などの各業界に向けて開発 IBM の人工知能マシン Watson の分析技術 cognitive computing 認知計算で「つぶやき」データを分析してビジネスに応用するという．

クイズ王に勝った人工知能コンピューター IBM Watson は，2011 年 2 月 14 日～16 日の 3 日間，アメリカ合衆国の人気クイズ番組「ジョパディ！Jeopardy!」でクイズ王の人間と戦った．行われたクイズ王対決の最終的な成績は，IBM の Supercomputer Watson が 7 万 7147 ドル，クイズ王のケン・ジェニングス氏は 2 万 4000 ドルで，ブラッド・ラッター氏は 2 万 1600 ドルだった（図 5）．

図５　人工知能コンピューターWatson

6.3　ビッグデータの未来

これからのビッグデータ活用が変える未来像を観ていくことにしよう．

（1）ビッグデータが変える医療

NHK スペシャル“新たな潮流　医療ビッグデータ”（2014.11.02,21:00-21:50）が放映され，医療への有功活用事例が紹介された．

1)病気を「予知」，命を守れ（US Rhode Island 州）では，オンタリオ工科大学教授のキャサリン・マクグレゴーさんが，新生児集中治療室の未熟児を，感染症を予知して救った，

2)最先端！ビッグデータ病院（済生会熊本病院）では，患者にセンサーを付けて，300 項目のデータを収集し，早く退院と相関のある 3 大要素（食事再会の早さ，点滴の期間の短さ，痛みの度合いの少なさ）を解明し，リハビリを早く始め，入院期間を半分に短縮した．

3)町ぐるみで「ぜんそく」激減（US Kentucky 州）では，吸入器を使って，発作の起きた原因を解析し，発作の回数が半減した．発作のポットスポットを調査し，原因を調べるための大気調査を開始，「南西の風」を解明した．

このように，少子高齢化社会で医療コスト

の負担を軽減するための「予防医療の推進」するため，電子カルテの標準化，徹底したIT化を進め，感染症の予測，伝染病からの被害を最小限にすること．また，DNA の解析から衛生管理を徹底し，不老長寿へ向かう．

（2）ビッグデータが変える交通インフラ

米国自動車保険業界は,テレマティクス(遠隔で走行位置や速度などのデータを収集するシステム)を利用し,走行データを分析して,運転状況を保険料に反映している．

トヨタやホンダの活用例に始まり，Google が推進する自動運転システムや，物流業界での効率的輸送システムでコストを削減し，渋滞情報，危険回避情報の提供で，円滑なトラフィックが確保できるようになる．

（3）ビッグデータが変えるその他の未来

・ビッグデータがブラック企業･行政を駆逐

・ビッグデータが変える「データ都市戦略」

・ビッグデータが変えるエネルギー･･･Smart Meter の導入で光熱費の 30%のコスト削減.

・ビッグデータが変える教育･･･Tablet と eBook, e-Learning, MOOCs (Massive Open Online Course：巨大でオープンなオンラインの授業)，ネット大学などで，場所，時間，金銭，年齢，学力，学校の定員などのような条件に縛られることなく，世界トップクラスの大学の講義や，著名な学者による講義などを試聴することができ，学生の訪問履歴，成績等の膨大なビッグデータを収集,分析して,授業に反映させ,授業評価が行われる.また,生き残りをかける大学経営に，教育 IR 戦略など，ビッグデータ解析が不可欠になる．

･ビッグデータ社会の新しい専門家･･･データを収集する会社データ･アグリゲーターData Aggregator，益々ニーズが高まる DS（Data Scientist)，Big Data を調査・分析し，公正に評価するアルゴリズミスト Algorithmist や Chief Analytics Officer 達が，センサーだらけの IoT（Internet of Things）の普及に伴って，闊歩する時代が来るだろう[10].

謝辞：本原稿のベースとなった講演「最近のビッグデータ活用事情」の機会を与えてくれた「日本技術士会」北陸本部富山県支部に感謝の意を表する．

参考文献と参照ウェブサイト等

［1］“格差広げるビッグデータ 100”，日経コンピューター，日経 BP 社，No.865，2014.07.24，28-53，2014.

［2］ビッグデータ・ビジネス，鈴木良介著，日経文庫，2012.10.15，ISBN978-4-532-11268-4，￥860+TAX

［3］ビッグデータの正体－情報の産業革命が世界のすべてを変える－，ビクター・マイヤー=ショーンベルガー，ケネス・クキエ著，斉藤栄一郎訳，講談社，2013.05.20，ISNB978-4-06-218061-0, ¥1,800+TAX

［4］ビッグデータの衝撃－巨大なデータが戦略を決める－，城田真琴，東洋経済，2012.07.12，ISBN978-4-492-58096-7，¥1,800+TAX

［5］ビッグデータ早わかり A Quick Illustrated Guide to Big Data，大河原克行著，中経出版，2013.01.29，ISBN978-4-8061-4620-9，￥1,500+TAX

［6］ビッグデータの覇者たち，海部美知著，講談社現代新書，2013.12.03，ISBN978-4-06-288203-3,¥760+TAX

［7］進撃のビッグデータ，牧野武文著，マイナビ新書，2014.06.30，ISBN978-4-8399-4961-7，￥850+TAX

［8］データサイエンティスト協会：http://www.datascientist.or.jp，スキル定義：http://prtimes.jp/main/html/rd/p/000000005.000007312.html（2015.1.30，確認）

［9］“ビッグデータは人工知能に任せた！”,，日経コンピューター，日経 BP 社，No.870，2014.07.24，22-39，2014.

［10］データ・アナリティクス 3.0　ビッグデータ超先進企業の挑戦，トーマス.H．ダベンポート著，小林啓倫訳，日経 BP 社，2014.5.7，ISBN978-4-8222-5013-3，￥2,000+TAX